ISSN: 2640-4133

Research Article

Advances in Bioengineering & Biomedical Science Research

Early Prediction of Lupus Disease: A Study on the Variations of Decision Tree Models

Jagjiven Kaur Jasber Singh [1*], Raja Rajeswari Ponnusamy[2], Elaine Chan Wan Ling[3] and Lim Sern Chin[4]

[1,2] School of Computing, Asia Pacific University, Malaysia,

[3] international Medical University, Malaysia

[4] Universiti Teknologi Mara, Malaysia

[*] Corresponding author
Jagjiven Kaur, School of Computing, Asia Pacific University, Kuala Lumpur, Malaysia.

Submitted: 13 Oct 2022; **Accepted**: 21 Oct 2022; **Published**: 29 Nov 2022

Citation: Singh, J. K. J., Ponnusamy, R. R., Ling, E. C. W., Chin, L.S. (2022). Early Prediction of Lupus Disease: A Study on the Variations of Decision Tree Models. *Adv Bioeng Biomed Sci Res*, 5(4), 204-215.

Abstract

Systematic Lupus Erythematosus (SLE) is an irreversible autoimmune disease that has seen to bring a lot of negative effect on the human body. It has become a very challenging task in predicting the prevalence of Lupus in patients. It has slowly gained popularity among many researchers to study the prevalence of this disease and developing prediction models that not only study the prevalence of the disease but is also able to predict suitable dosage requirements, treatment effectiveness and the severity of the disease in patients. All of these is usually done with medical records or clinical data that has different attributes related and significant to the analysis done. With the advancement in machine learning models and ensemble techniques, accurate prediction models have been developed. However, these models are not able to explain the significant contributing factors as well as correctly classify the severity of the disease. Decision Tree Classifier, Random Forest Classifier and Extreme Gradient Boosting (XGBoost) are the models that will be used in this paper to predict the early prevalence to Lupus Disease in patients using clinical records. The most significant factors affecting Systematic Lupus Erythematosus (SLE) will then be identified to aid medical practitioners to take suitable preventive measures that can manage the complications that arise from the disease. Hence, this paper aims to assess the performance of tree models by performing several experiments on the hyper parameters to develop a more accurate model that is able to classify Lupus Disease in patients in the early stages. Findings revealed that the best model was the Random Forest Classifier with parameter tuning. The most significant factor that affected the presence of Lupus Disease in patients was identified as the Ethnicity and the Renal Outcome or the kidney function of the patients.

Keywords: Systematic Lupus Erythematosus (SLE), Machine Learning (ML), Ensemble Techniques, Decision Tree Classifier, Random Forest Classifier, Extreme Gradient Boosting (XGBoost) Classifier, Clinical Data.

Introduction

Introduction to Lupus Disease

Systematic Lupus Erythematosus (SLE) is an autoimmune disease in the human body that is seen to bring a lot of negative impacts to individuals. The complexity in the disease prediction and the severity of its impact to patients makes it a challenging task to perform any type of analysis on the disease prevalence. According to Gergianaki and Bertsias, SLE is seen to cause damage to multiple organs in the human body with different level of severities and in some cases can even cause death [1]. SLE as according to Maidhof and Hilas, can be described as a multisystemic inflammation that is caused by the improper function of the immune system [2]. There are four different categories that SLE can be divided into which are shown below.

✓ **Neonatal and Pediatric Lupus Erythematosus (NLE)** – rare form of Lupus observed in neonates that is passed to them through their mother, affecting only 1% of the population.

✓ **Discoid Lupus Erythematosus (DLE)** – causes chronic scarring and dermatological skin sensitivities which may progress to SLE or develop in patients with SLE, more prevalent in women between the ages of 20 to 40 years old.

✓ **Drug-Induced Lupus (DIL)** – usually cause due to the exposure to certain medications that trigger an autoimmune response affecting various organs.

✓ **Systemic Lupus Erythematosus (SLE)** – most common type of lupus, affects 20 to 150 individuals out of 100,000, affects multi-organs systems, commonly found in young females but can be seen present in anyone either male or female.

Based on the different types of Lupus Disease, this paper focuses on the fourth type of Lupus which is the Systemic Lupus Erythematosus (SLE). Since it is the most commonly found among individuals, it is important to understand the effects of this disease on the human body and what can be done to prevent the development of the disease. In an article by Dorner and Furie, it was found that the death rate has declined, and patients now diagnosed with Lupus are able to live up to 15 years. Based on a report in 2017, it was found that the ancestral line, the race of an individual and the ethnicity of an individual plays a huge role in the manifestations of Lupus in individuals [3].

It is seen that Lupus Disease is becoming a growing worry in the medical field. Although there has been tremendous improvement that is observed in the prediction and the management of the disease, there still exists a gap that is the inability to predict the disease in an early stage. As such this will be the focus area of this research in developing a model that can perform early prediction of Lupus in patients based on their medical records in order to take preventive measures that can help lessen the sufferings of patients.

Background and Problem Statement

According to the World Bank, Malaysia currently has a population of 31.95 million individuals and is widely populated with individuals of different ethnicities inclusive of Malay, Chinese, Indian, Sikhs and even European cultural influences. The prevalence of SLE among Malaysians is worrying. In a study by Chai, Phipps and Chua, genetic susceptibility, environmental and hormonal factors are seen to be the most influential when it comes to the diagnosis of Lupus in patients [4]. As compared to Malay and Indian individuals, Chinese individuals were more affected by this disease. When developing models to predict the prevalence of Lupus Disease in patients, it is important to identify the patterns of that disease in the country of interest.

In a study by Yeap et al., the mortality rates of patients with Lupus were seen to be out of 494 patients, there was a 20.2% percent mortality rate, and the highest cause came from infection which affected 30% of the patients [5]. However, it was noted that 15% of these patients died due to renal problems. Lee et al. had suggested that from 2016 to 2019, there was a significant increase in the number of patients admitted with SLE while in the year 2018 to 2019 there was a 65.5% increase that was observed [6]. This raises concerns since the numbers are increasing. It was also found that the in-house mortality rate of patients with SLE was seen to have experienced an increase (10.7%) as compared to the previous years (3.7 – 5.4%). Flare and infection were the leading cause of death which was similar to the results that was obtained from the study by Yeap et al. This indicates that most of the deaths caused by SLE are not changing much although there have been improvements that are being done on the prediction of disease in this field of study [5].

Over the years, machine learning models and prediction techniques were developed. It was found that these techniques were useful for disease prediction. These machine learning techniques included models such as Artificial Neural Network (ANN) and the Naïve Bayes Classifiers which were used to predict the occurrence of Lupus Disease in patients as well as the severity of illness in patients with SLE. These models were proven to provide more accurate prediction results as compared to previously adopted methods [7]. Adamichou et al. suggested that Machine Learning tools are more commonly adopted to mimic the "Medical Reasoning" capability of humans in the medical field and is able to handle complex tasks effectively [8].

Over the years, it has been noticed that many researchers' have analyzed the trends and patterns that exists within the scope of Lupus Disease in patients. The mortality rates associated with SLE, contributing factors, damage prediction, analysis of remission therapy as well as an overview of the disease spread are some of the areas in which predictions have been done by previous studies [9-15]. These researchers used clinical records of patients along with the SLE symptoms that are stated by SLICC.

Machine learning models have been commonly used to perform analysis on Lupus Disease. However, deep learning models were not as preferred and did not show desired results as compared to machine learning classification and tree models. On the contrary, these researches stopped to the extent of just the prediction results from these models. The significant factors that contribute to the disease were not very frequently discussed. Limited study has been done on the application of application of the variation of tree models on the prediction of the significant factors affecting Lupus Disease in patients. Despite the development of many prediction models, these models were still unable to fully explain the symptoms of this disease, which does not give medical workers much confidence when handling patients with this disease due to the unpredictability of the disease. Past models lacked the ability to describe the future trends and the patterns that are experienced if there is no proper measure in place to control the spread of this disease.

Although all the journals referred to are able to develop a prediction model, the analysis done did not clearly state the outcome and prediction accuracies of the models developed, making it a challenge to evaluate the past performance of these techniques. Hence, there is a need to study the benefits of the application of variations of decision tree models in machine learning in the medical field of study. In the light of this, this study aims to perform a study on the decision tree models using previously applied decision tree models of high accuracy to develop a better and improved prediction model for Lupus Disease. This analysis would also use the predicted results to identify the factors that are most significant in the prevalence of Lupus in patients

Lupus Disease and Application of Machine Learning Techniques

Impacts of Lupus Disease

Being an incurable disease, Lupus has many negative effects the daily life of individuals suffering from it. It is seen to affect areas such as work, finances, school, relationships and even family relationships. It takes a toll on an individual making it difficult to carry on with regular daily tasks. Lupus comes with many side effects inclusive of but not limited to chronic pain, severe fatigue as well as inability to work to their best abilities due to the complications faced from the disease. It attacks not only adults but also known to affect children of young age. In a study by Macejova, Zarikova and Oetterova, SLE was found to have a significant effect on the professional activities that are carried out by individuals [16]. Around 39% of individuals that were a part of the study had stated that they were not able to perform their jobs well due to the disease hence they needed to change what they worked as. SLE affected patients even during periods of inactivity causing them to be fatigued, pain and even the inability to perform physical activities.

Several studies have also indicated that patients with Lupus also have been known to affect the performance of individuals in their work life as well. Clinical symptoms are also significant in identifying the reduction in productivity that patients experience due to Lupus. The treatment that are done to reduce the effects of this disease are seen to negatively affect the patient's life in almost 70% of the recorded cases. In a few studies that were done in the United States and United Kingdom had found that a significant percentage of individuals in the work force that experienced disabilities were those that were diagnosed with Lupus which accounted for about 20% of the work force [17, 18].

In the studies by used group of patients that self-diagnosed the severity of their symptoms [19, 20]. It was found that similar negative impacts were found to affect these individuals' lives making it hard for them to perform well at work. Majority of the cases affected women (93.1%) who were below the age of 50 (86.7%). 27.7% of these individuals were also seen to make a change in their careers due to the complications from Lupus Disease.

Based on these papers, it can be seen that Lupus has many negative impacts on an individual's life. It is also found that a lot of these symptoms are more prone to affecting women, indicating that gender plays a significant role in the spread of the disease. Patients with Lupus are also suffering from extreme fatigue which is directly affecting their performance at work. The drop-in work performance is an important issue that needs to be investigated. This in the long run could negatively impact the country's economic growth if the spread of the disease is not control through preventive measures. The quality of life will also decline making it difficult for patients to carry out daily activities. It is important for the top management at working facilities to create a more conducive environment for individuals with Lupus to work and maintain similar productivity levels.

Machine Learning and Ensemble Methods for Early Prediction of Diseases

Machine learning and ensemble models are effective when past data is being analysed using the past data to identify patterns that exists. The machine learning algorithms use the identified patterns and learns them to create value of the data that is being analysed. The analysis of these models allows the implementation of this combination of models not only in the medical domain but also other domains such as the financial markets. Large amounts of data can be fed into these models. Machine learning models are commonly used to perform forecasting and analysis in many different domains. This study aims to use the combination of different machine learning models to perform early prediction of Lupus Disease.

There have been many research papers that apply machine learning algorithms to perform predictions and are identified as effective for early prediction of disease with high accuracy. Ibrahim and Abdulazeez stated in their research that machine learning models are very useful in the medical sector and it can help improve the diagnostics of diseases among patients [21]. Some of the commonly applied models in the medical sector includes Naïve Bayes Classifier, K-Mean Clustering, Decision Tree and Deep Learning models. Using this as a reference for this study, machine learning methods such as the Decision Tree classifier will be used to predict the occurrence of Lupus in patients using their medical records.

In a study by Panicker and P, a brief review was done on the cardiovascular disease prediction using machine learning techniques [22]. It was found that with early detection of this disease in patients, the cost of medical care can be reduced, and a more reliable and accurate prediction model can be developed for the medical sector. This paper found that machine learning models as well as ensemble models were useful with data such as electronic health records and CT images. Support Vector Machines (SVM), neural networks and ensemble techniques were the best models whereby these models had accuracies of prediction above 95%. Yekkala, Dixit and Jabbar performed a similar study on the prediction of heart disease using ensemble techniques and swarm optimizations to improve the accuracy of the model [23. It was found that out of all the models used, the bagged tree models were one that achieved the highest accuracy.

Prediction analysis was also found to be done on the analysis of diabetes. Multiple machine learning models as well as ensemble techniques were also applied and found to have high accuracies of prediction. In the prediction of Diabetes and Diabetes Retinopathy by, machine learning models and ensemble models were used to perform the analysis [24-26]. When the comparative analysis was done on the prediction accuracies of these models, the ensemble methods were shown to outperform the individual machine learning models and similar to the analysis by other researcher, Naïve Bayes Classifier was the second best after the ensemble methods.

According to the analysis done, machine learning methods were also applied in the analysis of other chronic diseases such as Cancer and Parkinson. Kumar et al. performed an early prediction of cancer disease in patients using machine learning models [27]. Having similar views as the analysis by, the ensemble method used was Random Forest and was found to be the model that gave the highest prediction accuracy of 93% [26]. The paper by Lu et al. found that with the adoption of an ensemble technique known as Voting strategy, the accuracy of prediction can be tremendously improved in the prediction of cervical cancer in patients [28]. Tiwari et al. performed an early prediction of Parkinson Disease using machine learning and deep learning approaches and found that although there is no cure for the disease, the XGBoost model was able to predict the occurrence of this disease with an accuracy of 95% which outperform all the other models used [29].

With the prediction of Lupus Disease, there have also been a few researches that performed prediction using machine learning and ensemble methods. Being an autoimmune disease in the human body, it can be difficult to accurately have an early prediction of the disease to avoid further and more risky damage to the organs. According to Stafford et al., the integration of AI technologies and machine learning models is able to allow the stratification of patients by different categories of severity as well as performing diagnosis, management of the disease, evaluating the different response of patients to the treatment provided and the risk that is associated to it [30]. It was found that the application of machine learning models can provide a better understanding of the development and the progression of a disease which was successfully explained by machine learning and deep learning models that can handle large amounts of data [31-33]. With appropriate optimization techniques and enhancement to the hyper parameters applied to the models, these prediction accuracies can be further improved and is able to transform the diagnosis of Lupus in healthcare institutes.

Performance Evaluation Techniques

Based on the analysis done, it is important to analyse the performance of these models using appropriate performance evaluation tools. This ensures that the prediction capability can be captured and if it is not of good quality, it can be further improved. Being able to analyse the capability of these models, will ensure that an efficient model can be developed. Referring to the researches done on the prediction of disease, the Accuracy, Precision, Recall and Confusion Matrix were some of the most popular measures of performance that were used. In relation to this, this study will use the aforementioned methods to evaluate the performance of the models that are being used in this analysis to make an early prediction of Lupus Disease in patients using the medical records that is available.

Discussion & Summary

The literature review showed that in the field of diseases prediction, there are many machine learning models that can be used to perform the analysis. It was found that Ensemble models were of much higher accuracies as compared to standalone machine learning models. The Decision tree models were also an area that showed significant prediction accuracies when it involves diseases and patients' medical records. In references to this, the models that will be applied in this study would include a range of ensemble techniques and machine learning models to perform an early prediction of Lupus Disease. The models include the Decision Tree Classifier, Random Forest Classifier and the Extreme Gradient Boosting (XGBoost) ensemble model. The explanation of these models will be explained in the following chapters.

Based on the analysis done, it was found that although quite a number of papers were written on the development of prediction models for diseases, there were fewer studies of the application of machine learning techniques on the early prediction of Lupus Disease in patients. The predictive capabilities of these models were seen to be very beneficial to the medical sector and can add a lot of value to the diagnosis that is made.

It was found that Decision Tree models were not new techniques. It had been widely used in the medical sector. However, based on the papers, it was found that there were not many studies that evaluated the performance of the different variations of decision tree models on the prediction of lupus disease. The papers also did not clearly identify the significant factors that affect lupus disease in patients. Moreover, the model selection criteria were also discussed in this chapter that will be used to evaluate the prediction accuracies of the models developed. Thus, this paper will aim to bridge this gap that exists in the prediction of lupus disease in patients.

Research Methodology

Research Approach

This research is a combination of the quantitative and qualitative approach whereby it uses secondary dataset that was obtained from the International Medical University (IMU) for the purpose of early prediction of Lupus Disease in patients. Since the prediction of Lupus Disease in patients' is done with the intention to support the aim and objectives of this research, a positivist approach is taken.

There are five stages that this research will follow to conclude the research problem that has been stated. Stage 1 one this research will be the data collection stage where all the necessary data needed for analysis is collected. Once this is done, in stage 2, the data is then explored to understand the patterns that exists within the data, missing values as well as variables that may transformation. The third stage is when the data pre-processing is carried out. Any irregularity that was observed in stage 2 will be adjusted and modified in this stage. The cleaned data is then fed into the appropriate tool to develop models which will be evaluated using simulations and performance evaluation techniques. The final models will then be tested on the testing dataset.

There will be three models that will be applied to the dataset to perform the early prediction of Lupus Disease in patients. The predictive modelling techniques that will be applied are the Decision Tree Classifier, Random Forest Classifier and Extreme Gradient Boosting (XGBoost). Using Accuracy, Precision, Recall and the Confusion Matrix, the best model will be selected. The flowcharts in Figure 1 and 2 show the research framework and the illustration of the methodology applied in this research.

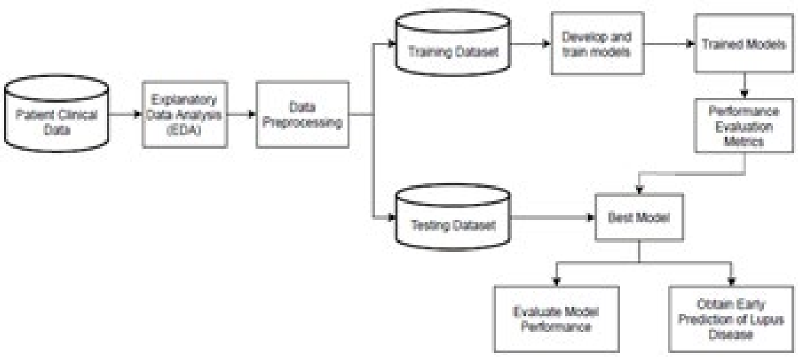
Patient Clinical Data
Explanatory Data Analysis (EDA)
Data Preprocessing
Training Dataset
Develop and train models
Trained Models
Performance Evaluation Metrics
Testing Dataset
Best Model
Evaluate Model Performance
Obtain Early Prediction of Lupus Disease

Figure 1: Research Mechanism

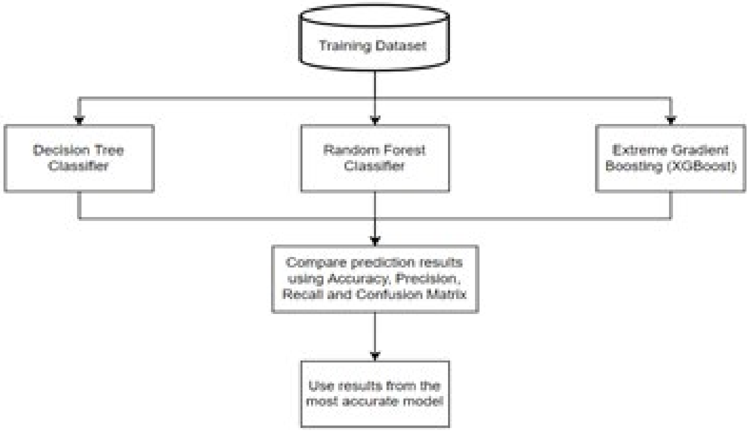
Training Dataset
Decision Tree Classifier
Random Forest Classifier
Extreme Gradient Boosting (XGBoost)
Compare prediction results using Accuracy, Precision, Recall and Confusion Matrix
Use results from the most accurate model

Figure 2: Proposed Methodology

Description of Dataset

This dataset consists of 141 patients' clinical records and it has 66 different attributes that describe the medical history and symptoms of a patient. The attributes of the dataset have demographic information of the patients, clinical features at diagnosis, Systemic Lupus International Collaborating Clinics (SLICC) Systematic Lupus Erythematosus (SLE) criteria, Selena SLEDAI (Systemic Lupus Erythematosus Disease Activity Index) Score, Disease Progression over the period of 10 years, Disease Damage over the period of 10 years, Renal Disease Treatment, Non-renal flares, and the current status of the patients. All these variables are classified using numerical values for different levels in each variable.

Description of Models Adopted

Decision Tree Classifier

The Decision Tree model is a commonly applied model that is used for both classification and regression tasks. It is one of the simplest yet powerful machine learning model that proves to be very beneficial. One great application of this model is that is can be applied even to non-linear data and usually it is known the give the best outcome. The algorithm evaluates the dataset and use the attributes to come up with different solutions to the problems. The decision trees are also fairly easy to interpret thus it is a suitable model to be applied for the classification and prediction of lupus disease in patients. According to Figure 3 below, the general decision tree works in a similar manner (Roy 2020) [34].

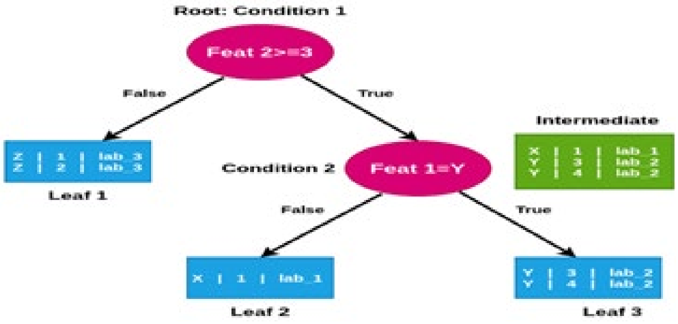

Figure 3: Decision making process flow

Random Forest Classifier

The Random Forest Classifier is another variation of the decision tree model and is known as an ensemble technique. It is a classification algorithm that uses several decision trees to come up with a solution. The best tree algorithm will be selected and used for the predictions that are being done. The randomness that the model provides allow the reduction of model biasness. This model is able to prevent the overfitting of data in the model. It has been seen to be used in many different fields such as the medical field, stock market and even the banking field. According to You (2019), the random forest works as indicated in Figure 4 below. Several trees are plotted and the best is selected to perform the prediction.

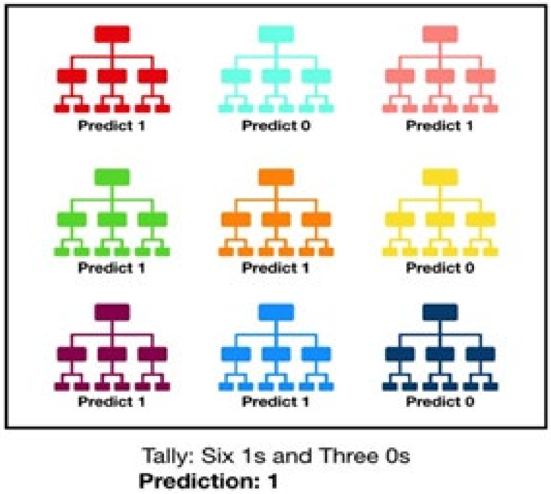

Figure 4: Process Flow of Random Forest Classifier

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a tree-based model which is a commonly applied technique in classification problems. It is an extensive of the Gradient Boosting model. This model is suitable for both regression and classification tasks however it works better when used for classification problems. Referring to the article by Dobilas, Gradient Boosting and XGBoost both make use of decision trees as base estimators and formulated several trees in the modelling process [35]. The result of the classification is done by taking the average prediction of all the formulated decision trees.

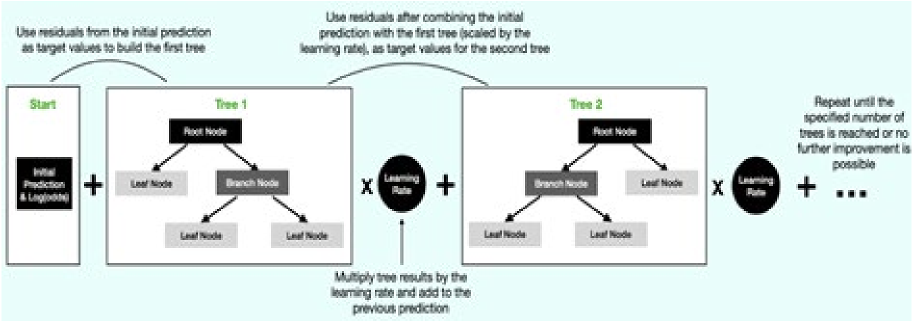
Use residuals from the initial prediction as target values to build the first tree
Use residuals after combining the initial prediction with the first tree (scaled by the learning rate), as target values for the second tree
Start
Initial Prediction & Log(odds)
+
Tree 1
Root Node
Leaf Node
Branch Node
Leaf Node
Leaf Node
X
Learning Rate
+
Tree 2
Root Node
Branch Node
Leaf Node
Leaf Node
Leaf Node
X
Learning Rate
+ ...
Repeat until the specified number of trees is reached or no further improvement is possible
Multiply tree results by the learning rate and add to the previous prediction

Figure 5: Process Flow of Extreme Gradient Boosting (XGBoost)

Figure 5 above shows the basic process flow of how a prediction is made using the XGBoost model. It can be seen that the trees are scaled using the individual learning rates to form the consecutive tree until the specified number of trees is reached. The residual values are used to form the decision trees rather than the actual class labels creating regression trees to solve classification problems.

When developing an XGBoost model, it uses the Similarity Index and gain to identify the best tree. The formula that is used to form the tree is only slightly different from the Gradient Boosting where there is an additional lambda term in the formula. This is a regularization term where it can be used to reduce the influence of smaller leaves in the tree by increasing the lambda value. The gain is then calculated for these trees to evaluate the best node split. The formulas used for computation of these trees are shown below in Equation (1) and (2).

“Similarity Score “ $(\sum_{i=1}^{n} \text{Residual}_i)^2 / (\sum_{i=1}^{n} [\text{Previous Probability}_i \times (1-\text{Previous Probability}_i)] + \lambda)$ (1)

Gain Left $\text{Leaf}_{\text{Similarity}}$+Right $\text{Leaf}_{\text{Similarity}}$-$\text{Root}_{\text{Similarity}}$ (2)

λ = Regularization Parameter

Where;

Residual = Observed/Actual Value – Predicted Value

Previous Probability = The probability of an event calculated in

the previous step
Note: Probability is always 0.5 for the first tree

Several disease predictions papers have also opted for the XGBoost modelling technique compared to the regular Gradient Boosting model. The prediction of Heart Disease, Liver Disease, Type 2 Diabetes Risk, and the prediction of Breast Cancer in patients were some of the studies that found this method to have high prediction accuracies [36-38]. The accuracies of these modelling techniques were found to be above 85% and majority of them were 90% percent and above. One of the papers by Murty and Kumar showed that with the optimization of L2 regularization, Logistic loss function, learning rate and the number of estimators that are used in the model development gave them an accuracy of 99% which is the highest recorded in the previously done studies [38]. Budoliya, Shrivastava and Sharma used the Bayesian Optimization on their model and managed to get an 85% accuracy on the training dataset and a 91.8% accuracy on the testing dataset [36].

Similarly, there were also several papers that employed the XGBoost modelling technique on the prediction of Lupus Diseases in patients and most of the accuracies recorded were above 75%, which is considered a good accuracy score for a disease prediction model . This paper aims to use the researchers' previously done to further optimize models applied in the prediction of Lupus Diseases in patients using the clinical data that has been recorded [33, 39].

Performance Evaluation Techniques

When the fitting and predictions for every model is completed, the performance of the individual models must be assessed to understand the accuracy of prediction of each model. Four performance evaluation techniques will be used to evaluate the performance inclusive of Accuracy, Precision, Recall and Confusion Matrix. Performing this evaluation on the models will assist in identifying the best model as well as provide better results that fulfil the objectives of this research.

Data Analysis and Implementation

Model Development

With the intention to develop a suitable predictive modelling for the purpose of early prediction of lupus disease in patients based off the medical symptoms that are exhibited by them, six different experiments were carried out. There were two experiments for each model, where the first experiment is the application of the base model while the second experiment is further enhanced with parameter tuning. The dataset that was used for all the models were similar. Once the models were fitted onto the dataset and the prediction was performed, the overall model performance and the results were put together for compilation as well as thoroughly analysed to identify the best model that fulfils the aim of this research. The six experiments that were carried out were inclusive of three tree models which are the Decision Tree Classifier, Random Forest Classifier and the Extreme Gradient Boosting Classifier (XGBoost). The base model were developed using the pre-set parameters while the second model for each technique is adjusted. Each experiment was trained using the training dataset and the testing dataset was used to perform prediction in order to test how well the model would perform on a real dataset for the classification of disease.

Decision Tree

The Decision Tree model is a commonly sourced tree model that can be useful for the classification task. There were two experiments that were created for this model where one is a base model while the other is a model with parameter tuning.

Experiment 1: Decision Tree Classifier

The fitted model is used on the test dataset in order to come up with predictions which were compared to the actual test values. With visualization techniques, it was found that although there were quite a few correct classifications, there is still a bigger number of cases that were wrongly classified. Metrics such as "accuracy", "precision", "recall", and "confusion matrix" were computed. As observed, the accuracy is very low for this model, which is only 39.53%. It does not indicate a very good fit. This could be due to the imbalances that were found in the target variable or it could be due to the lack of data to train the model. When the precision and recall is taken into account, it can be seen that the highest precision obtained by the model is 0.556 and it goes as low as 0.000. The "0.000" could most probably be due to a very small sample for "Class 5". The recall on the other hand has its top value as 0.600 to as low as 0.000. Overall, the model seems to be poor in the task of classifying patients into categories.

Figure 6: Output – Decision tree (Experiment 1)

The decision tree in Figure 6 above looks very complex because all the variables are considered. This could become very complicated and may prove harder to analyse. It can be identified that "Out_R", "Ethnic" and "PD_mth" were the most significant factors.

Experiment 2: Decision Tree Classifier with Parameter Tuning

The prediction results that were obtained from the second model were found to have many instances where the model was not able to correctly identify which class the patients belonged to. Thus, we can say that the model is not accurately predicting patients.

The accuracy of this model is seen to be 58.14%. This value is fairly acceptable, as a high accuracy would indicate that the model is performing better. However, to evaluate the performance, alone is not enough. In that case, the "precision" and "recall" is computed to further evaluate the model. The precision indicates the ability of the model to identify the true positives among all the other positives, while recall is the number of true positives from the entire data. The precision for this model has a highest value of 0.600 and a lower value of 0.400 while the recall has a value of 0.920 and 0.182. These values show that although the accuracy of the model is important, precision and recall prove to be more beneficial in evaluating the performance of a model.

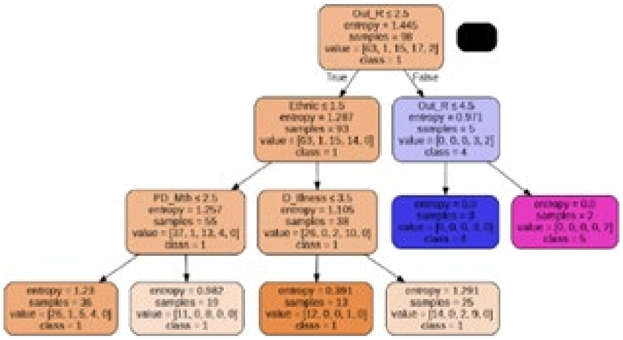

Figure 7: Output – Decision tree with parameter tuning (Experiment 2)

The decision tree in Figure 7 above is fairly easier to read and understand as compared to the first visual that was developed. The tree indicates that the top and most significant variable is the "Out_R" variable as this variable is the variable that indicates the renal status of the patients.

Random Forest

Experiment 3: Random Forest Classifier

The prediction results that were found for the first experiment using this model indicated that quite a big chunk of the predictions that were made were not correctly predicted. This indicates that the model did not fit the data very well.

The accuracy of the model is 55.81% which is not very high, but it is acceptable. The precision of the model shows that it has the capability of correctly identifying the "Current Status" of patients up to 58.5% and the recall of the model was seen to be 96% which indicates that among the entire data, the model can correctly predict the right status of lupus in patients.

The random forest tree was plotted once predictions were made, however, since this is the base model and no specific adjustments were made to reduce the size of the parameters that were taken into account, the Random Forest tree developed was very complex and it proved to be a challenge in understanding the output. It was found that "Ethnic" and "C_nephrotic" were the top attributes.

Experiment 4: Random Forest Classifier with Parameter Tuning

Feature importance analysis is an importance step in analyzing the attributes that are found in the dataset. Doing so will allow the selection of the best attributes that would be suitable to be fitted to the model. From this output, six of the variables will be used to create a new dataset to fit to the model which are "D_Illness", "Ethnic", "PD_Mth", "D_Age", "Out_R" and "C_hypert".

From the predictions, it can be seen that there is a very big difference in the prediction and the actual results. All the predicted results were from the status of "1" while in the test dataset there was a fluctuation between the different classes. This indicate that the results are not very good as the model did not fit well. The accuracy of the model was seen to be 58.14%. This is not very different from all the other models however it does appear to be slightly higher than the rest. Using the precision as a guideline, it was seen to be about 58.1% while the recall was 100%. These results show that the model is capable of correctly predicting more than half of the correct classes while it has the capability to predict correctly among the entire datasets.

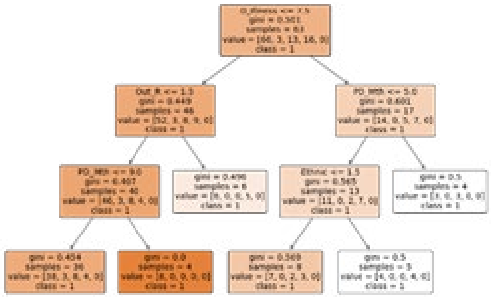

Figure 8: Output – Random forest tree

Based on the Random Forest tree above in Figure 8, it was found that "D_Illness", "Out_R" and "PD_Mth" were the most important features for analysis.

Extreme Gradient Boosting (XGBoost)

Experiment 5: XGBoost Classifier

The prediction outputs shows that a fair amount of the results was correctly identified however there were some cases where it was misclassified. The accuracy of the model was 48.84% which is very low. It shows that the model needs to be adjusted to fit the dataset better to achieve better results. Referring to the precision which is seen to be 58.8% and the recall was 80%. The prediction accuracy of the model in correctly identifying the status of the patients is good.

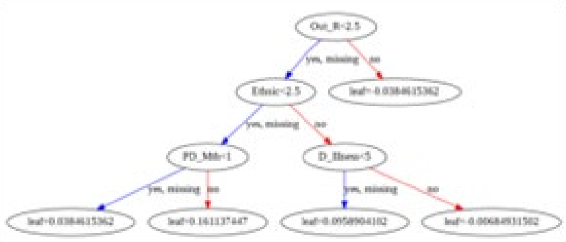

Figure 9: Output – Plot tree (XGBoost) (Part 1)

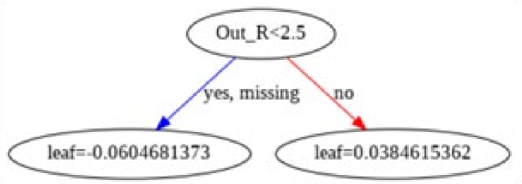

Figure 10: Output – Plot tree (XGBoost) (Part 2)

This visualization will explain the most important factors that are significant to the "Current Status" of the patients with lupus disease. Figure 9 and 10 shows the trees that were plotted from the experiment. According to the results, it was found that the most important attribute was "Out_R" and "Ethnic".

Experiment 6: XGBoost Classifier with Parameter Tuning

The prediction outputs for this experiment indicates that the model was able to classify the classes well at some points of the data but at some it showed errors. The accuracy of the model was found to be 55.81% which shows that it has significantly increased from the base model, but it still is not able to fully classify the "Current Status" of the patients correctly. The precision and recall of the model were also computed which was 59% and 92%, respectively. This indicates that the model was able to correctly classify more than half of the correct classes and most of the cases were correctly classified as a whole. The visualization of the XGBoost tree was plotted for better understanding as shown in Figure 11 and 12 below. It was found that the most important attribute was the "D_Illness" and "PD_Mth".

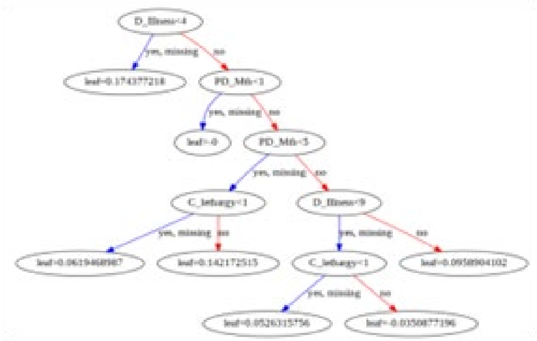

Figure 11: Output – Plot tree (XGBoost with parameter tuning) (Part 1)

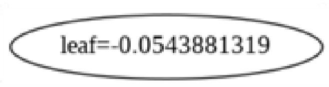

Figure 12: Output – Plot tree (XGBoost with parameter tuning) (Part 2)

Table 1 below shows the experiment summary and the accuracies for each of the models that have been experimented. It was found that the highest accuracy observed was 58.14%, which came from the Decision Tree Classifier with Parameter Tuning and the Random Forest Classifier with Parameter Tuning. . Both of the models were models that included hyper parameter tuning to ensure that the model fits better to the dataset.

Table 1: Experiments Summary and Results

No	Experiment	Accuracy
Decision Tree Classifier		
1	Decision Tree Classifier	39.53%
2	Decision Tree Classifier with Parameter Tuning	58.14%
Random Forest Classifier		
3	Random Forest Classifier	55.81%
4	Random Forest Classifier with Parameter Tuning	58.14%
Extreme Gradient Boosting (XGBoost) Classifier		
5	Extreme Gradient Boosting (XGBoost) Classifier	48.84%
6	Extreme Gradient Boosting (XGBoost) Classifier with Parameter Tuning	55.81%

In order to further narrow down to which is the best model for the purpose of classification and early prediction of lupus in patients, the precision and the recall scores indicate that the Random Forest Classifier is slightly better. However, to further evaluate this, there is a need for further analysis. Doing so will help identify and develop the best model that can be used by medical practitioners when making decision in the medical field.

Conclusion

To conclude, the tree models that were used within this study included Decision Tree Classifier, Random Forest Classifier and the Extreme Gradient Boosting (XGBoost) Classifier. Using the Google Colab environment, there were six different experiments that were carried out with the aim to identify the most significant factor affecting lupus disease and developing a classification model that is capable of handling the different symptoms to correctly classify the status of lupus in patients. Proper preprocessing steps were taken to ensure that the data that was being worked on was clean and would not cause any biases in the analysis.

The results that were achieved showed that there were two models which indicated similar accuracies of prediction and classification which was the Decision Tree Classifier and the Random Forest Classifier, in which both of the models were those that had been further enhanced by adjusting the hyper parameters of the model where the accuracy observed was 58.14%. If only one model had to be chosen, the Random Forest Classifier was the best by also taking into account the Precision and the Recall scores.

Based on the outputs, it was identified that the most significant factor that affects lupus to be triggered within patients are the "Ethnic" and the "Out_R". These factors are the Ethnicity of the patient as well as the Renal Outcome. This indicates that the individual background of the patient is important to be known as it had also been highlighted by a few researchers the most important factor that affects lupus in patients. Another possible area of focus is the kidney functionality. Renal treatment is used for patients when the kidney is not functioning up to its maximum capability, hence this treatment is carried out to help the functionality of the kidney. A bad performing kidney has a significant effect on the presence of lupus disease in patients and can be used as an identifier when evaluating patients to check for the presence of lupus.

During the timeline of this research, there were many challenges that were faced in the process of acquiring the dataset and the performing the different experiments. One of the biggest challenges was the time frame that was provided to complete the research. There was limited time and the scope of the research had to be thoroughly narrowed down to ensure that it could be completed within the stipulated time. Another challenge that was faced was with the dataset. Since this was a real dataset that was acquired from actual patients that were collected by the International Medical University (IMU), the data was very limited. There were only 141 patients' record that was available for use, which makes it difficult to get the best results when working with machine learning models. The dataset also showed issue with class imbalances which largely impacts the performance of the classification models. However, the dataset was used as it is to ensure that the results achieved are not tempered due to the adjustments made.

This project provides a baseline for future research to evaluate the different factors that affect lupus disease in patients. It was found that by adjusting the hyper parameters in a model proves to improve the performance of the model significantly. However, there were a few areas in which future work can be carried out. One area in which improvement and further work can be done is in the collection of the dataset. Future researchers can focus on collecting more data and reapply the models to evaluate if the performance can be further improved while taking this research as a guideline. Selecting the right hyper parameters is crucial in the application of any model especially tree models. Hence, future research can also be done on evaluating the hyper parameters that are suitable for classification problems by performing experiments. Additionally, the data splitting is also a crucial step in machine learning models. Thus, instead of selecting only one set split such as the 70:30 split ratio, researchers can study the application of cross-validation techniques in the selection of a suitable split for the train and test datasets. This may further enhance the performance of the models and improve the accuracies [40-45].

Declarations

Conflicts of interest. The authors declare that they have no conflict of interests.

References

1. Gergianaki, I., & Bertsias, G. (2018). Systemic lupus erythematosus in primary care: an update and practical messages for the general practitioner. Frontiers in medicine, 5, 161.
2. Maidhof, W., & Hilas, O. (2012). Lupus: an overview of the disease and management options. Pharmacy and Therapeutics, 37(4), 240.
3. Dörner, T., & Furie, R. (2019). Novel paradigms in systemic lupus erythematosus. The Lancet, 393(10188), 2344-2358.
4. Chai, H. C., Phipps, M. E., & Chua, K. H. (2012). Genetic risk factors of systemic lupus erythematosus in the Malaysian population: a minireview. Clinical and Developmental Immunology, 2012.
5. Yeap, S. S., Chow, S. K., Manivasagar, M., Veerapen, K., & Wang, F. (2001). Mortality patterns in Malaysian systemic lupus erythematosus patients. Medical Journal of Malaysia, 56(3), 308-312.
6. Lee, W. W. H., Cheong, Y. K., Teh, C. L., Wan, S. A., Chuah, S. L., & Singh, B. S. M. (2021). Impact of COVID-19 on hospitalization of patients with systemic lupus erythematosus (SLE). Clinical Rheumatology, 40(11), 4775-4777.
7. Ceccarelli, F., Sciandrone, M., Perricone, C., Galvan, G., Morelli, F., Vicente, L. N., ... & Conti, F. (2017). Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models. PLoS One, 12(3), e0174200.
8. Adamichou, C., Genitsaridi, I., Nikolopoulos, D., Nikoloudaki, M., Repa, A., Bortoluzzi, A., ... & Bertsias, G. K. (2021). Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. Annals of the rheumatic diseases, 80(6), 758-766.
9. Akbarian, M., Faezi, S. T., Gharibdoost, F., Shahram, F., Nadji, A., Jamshidi, A. R., ... & Davatchi, F. (2010). Systemic

lupus erythematosus in Iran: a study of 2280 patients over 33 years. International journal of rheumatic diseases, 13(4), 374-379.

10. Ginzler, E. M., Wallace, D. J., Merrill, J. T., Furie, R. A., Stohl, W., Chatham, W. W., ... & LBSL02/99 Study Group. (2014). Disease control and safety of belimumab plus standard therapy over 7 years in patients with systemic lupus erythematosus. The Journal of rheumatology, 41(2), 300-309.
11. Medina-Quiñones, C. V., Ramos-Merino, L., Ruiz-Sada, P., & Isenberg, D. (2016). Analysis of complete remission in systemic lupus erythematosus patients over a 32-year period. Arthritis care & research, 68(7), 981-987.
12. Ocampo-Piraquive, V., Nieto-Aristizábal, I., Cañas, C. A., & Tobón, G. J. (2018). Mortality in systemic lupus erythematosus: causes, predictors and interventions. Expert review of clinical immunology, 14(12), 1043-1053.
13. Stojan, G., & Petri, M. (2018). Epidemiology of systemic lupus erythematosus: an update. Current opinion in rheumatology, 30(2), 144.
14. Legge, A., Kirkland, S., Rockwood, K., Andreou, P., Bae, S. C., Gordon, C., ... & Hanly, J. G. (2020). Prediction of damage accrual in systemic lupus erythematosus using the Systemic Lupus International Collaborating Clinics Frailty Index. Arthritis & Rheumatology, 72(4), 658-666.
15. Restrepo-Escobar, M., Granda-Carvajal, P. A., Aguirre, D. C., Hernández-Zapata, J., Vásquez, G. M., & Jaimes, F. (2021). Predictive models of infection in patients with systemic lupus erythematosus: a systematic literature review. Lupus, 30(3), 421-430.
16. Macejová, Ž., Záriková, M., & Oetterová, M. (2013). Systemic lupus erythematosus--disease impact on patients. Central European Journal of Public Health, 21(3).
17. Agarwal, N., & Kumar, V. (2016). Burden of lupus on work: issues in the employment of individuals with lupus. Work, 55(2), 429-439.
18. Booth, S., Price, E., & Walker, E. (2018). Fluctuation, invisibility, fatigue–the barriers to maintaining employment with systemic lupus erythematosus: results of an online survey. Lupus, 27(14), 2284-2291.
19. Garris, C., Oglesby, A., Sulcs, E., & Lee, M. (2013). Impact of systemic lupus erythematosus on burden of illness and work productivity in the United States. Lupus, 22(10), 1077-1086.
20. Gordon, C., Isenberg, D., Lerstrøm, K., Norton, Y., Nikaï, E., Pushparajah, D. S., & Schneider, M. (2013). The substantial burden of systemic lupus erythematosus on the productivity and careers of patients: a European patient-driven online survey. Rheumatology, 52(12), 2292-2301.
21. Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends, 2(01), 10-19.
22. Panicker, S. (2020, June). Use of Machine Learning Techniques in Healthcare: A Brief Review of Cardiovascular Disease Classification. In 2nd International Conference on Communication & Information Processing (ICCIP).
23. Yekkala, I., Dixit, S., & Jabbar, M. A. (2017, August). Prediction of heart disease using ensemble learning and Particle Swarm Optimization. In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon) (pp. 691-698). IEEE.
24. Alehegn, M., Joshi, R., & Mulay, P. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics, 118(9), 871-878.
25. Reddy, G. T., Bhattacharya, S., Ramakrishnan, S. S., Chowdhary, C. L., Hakak, S., Kaluri, R., & Reddy, M. P. K. (2020, February). An ensemble based machine learning model for diabetic retinopathy classification. In 2020 international conference on emerging trends in information technology and engineering (ic-ETITE) (pp. 1-6). IEEE.
26. SK, S. (2017). A machine learning ensemble classifier for early prediction of diabetic retinopathy. Journal of Medical Systems, 41(12), 1-12.
27. Kumar, S., Kumar, H., Swarna, S. and Dutt, V., 2020. Early Diagnosis and Prediction of Recurrent Cancer Occurrence in a Patient Using Machine Learning. European Journal of Molecular & Clinical Medicine, [online] 7(7), pp.6785-6794.
28. Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Generation Computer Systems, 106, 199-205.
29. Tiwari, H., Shridhar, S. K., Patil, P. V., Sinchana, K. R., & Aishwarya, G. (2021). Early prediction of parkinson disease using machine learning and deep learning approaches. EasyChair Preprint, 4889, 1-14.
30. Stafford, I. S., Kellermann, M., Mossotto, E., Beattie, R. M., MacArthur, B. D., & Ennis, S. (2020). A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ digital medicine, 3(1), 1-11.
31. Simos, N. J., Dimitriadis, S. I., Kavroulakis, E., Manikis, G. C., Bertsias, G., Simos, P., ... & Papadaki, E. (2020). Quantitative identification of functional connectivity disturbances in neuropsychiatric lupus based on resting-state fMRI: a robust machine learning approach. Brain Sciences, 10(11), 777.
32. Jiang, M., Li, Y., Jiang, C., Zhao, L., Zhang, X., & Lipsky, P. E. (2021). Machine learning in rheumatic diseases. Clinical Reviews in Allergy & Immunology, 60(1), 96-110.
33. Chen, Y., Huang, S., Chen, T., Liang, D., Yang, J., Zeng, C., ... & Liu, Z. (2021). Machine learning for prediction and risk stratification of lupus nephritis renal flare. American Journal of Nephrology, 52(2), 152-160.
34. Roy, A. (2020). A dive into decision trees.
35. Dobilas, S. (2021). XGBoost: Extreme Gradient Boosting—How to Improve on Regular Gradient Boosting?
36. Budholiya, K., Shrivastava, S. K., & Sharma, V. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University-Computer and Information Sciences.
37. Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020, July). Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. In Healthcare (Vol. 8, No. 3,

p. 247). MDPI.

38. Murty, S. V., & Kumar, R. K. (2019). Accurate liver disease prediction with extreme gradient boosting. Int. J. Eng. Adv. Technol., 8(6), 2288-2295.
39. Dovgan, E., Gradišek, A., Luštrek, M., Uddin, M., Nursetyo, A. A., Annavarajula, S. K., ... & Syed-Abdul, S. (2020). Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients. Plos one, 15(6), e0233976.
40. Apte, A. (2018). 3 Ways to Load CSV files into Colab.
41. Dulhare, U. N. (2018). Prediction system for heart disease using Naive Bayes and particle swarm optimization. Biomedical Research, 29(12), 2646-2649.
42. Huang, Y., & Chung, A. (2020, October). Edge-variational graph convolutional networks for uncertainty-aware disease prediction. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 562-572). Springer, Cham.
43. Maliha, S. K., Ema, R. R., Ghosh, S. K., Ahmed, H., Mollick, M. R. J., & Islam, T. (2019, July). Cancer disease prediction using Naive Bayes, K-nearest neighbor and J48 algorithm. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.
44. Maltezos, S., & Georgakopoulou, A. (2021). Novel approach for Monte Carlo simulation of the new COVID-19 spread dynamics. Infection, Genetics and Evolution, 92, 104896.
45. Yiu, T. (2019). Understanding Random Forest.